PhyreRisk: A Dynamic Web Application to Bridge Genomics, Proteomics and 3D Structural Data to Guide Interpretation of Human Genetic Variants

Tochukwu C. Ofoegbu [1,†], Alessia David [1†], Lawrence A. Kelley [1], Stefans Mezulis [1], Suhail A. Islam [1], Sophia F. Mersmann [1], Léonie Strömich [1], Ilya A. Vakser [2], Richard S. Houlston [3] and Michael J.E. Sternberg [1]

1 - *Centre for Integrative Systems Biology and Bioinformatics,* Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK
2 - *Computational Biology Program and Department of Molecular Biosciences,* The University of Kansas, Lawrence, KS 66045, USA
3 - *Division of Genetics and Epidemiology,* The Institute of Cancer Research, London, SM2 5NG, UK

Correspondence to Alessia David: *Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK.* alessia.david09@imperial.ac.uk
https://doi.org/10.1016/j.jmb.2019.04.043

Abstract

PhyreRisk is an open-access, publicly accessible web application for interactively bridging genomic, proteomic and structural data facilitating the mapping of human variants onto protein structures. A major advance over other tools for sequence-structure variant mapping is that PhyreRisk provides information on 20,214 human canonical proteins and an additional 22,271 alternative protein sequences (isoforms). Specifically, PhyreRisk provides structural coverage (partial or complete) for 70% (14,035 of 20,214 canonical proteins) of the human proteome, by storing 18,874 experimental structures and 84,818 pre-built models of canonical proteins and their isoforms generated using our *in house* Phyre2. PhyreRisk reports 55,732 experimentally, multi-validated protein interactions from IntAct and 24,260 experimental structures of protein complexes.

Another major feature of PhyreRisk is that, rather than presenting a limited set of precomputed variant-structure mapping of known genetic variants, it allows the user to explore novel variants using, as input, genomic coordinates formats (Ensembl, VCF, reference SNP ID and HGVS notations) and Human Build GRCh37 and GRCh38. PhyreRisk also supports mapping variants using amino acid coordinates and searching for genes or proteins of interest.

PhyreRisk is designed to empower researchers to translate genetic data into protein structural information, thereby providing a more comprehensive appreciation of the functional impact of variants. PhyreRisk is freely available at http://phyrerisk.bc.ic.ac.uk

Introduction

The sheer scale of information from large genome projects such as the UK 100K Genomes Project has led to a formidable challenge of interpreting the functional consequences of genetic variants. Determining the effect of an amino acid substitution on protein structure is central to the interpreting genetic variants [1]. Experimental three-dimensional structures of proteins deposited in Protein Data Bank (PDB) [2], however, cover less than 20% of the human proteome. Homology modeling is a powerful tool for generating a 3D structure in the absence of experimental structure, and several homology modeling servers and databases are available. These include Rosetta [3], I-Tasser [4] and our *in house* program Phyre2 [5]. By means of homology modeling, structural coverage of the human proteome is significantly enhanced reaching 70% [6].

Current tools for mapping of variants using genomic coordinates onto PDB experimental structures are CRAVAT [7] and MuPIT [8]. Similarly, the G2S (Genome to Structure) server provides web APIs for the mapping of UniProt positions onto experimental protein structures

[9]. Protein structure resources focusing on genetic variant interpretation in the context of cancer genomics are Cosmic3D [10] and MOKCa [11] that allow mapping of variants reported in the Cosmic database onto experimental 3D structures. Interactome3D [12] and Interactome INSIDER [13] provide structural coverage of protein–protein complexes. Other resources, such as LS-SNP/PDB [14] and SNP2structure [15], have unfortunately not been maintained.

The tools that are now available for mapping variants to structures have two major limitations. First, they map to experimental structures but do not support the use of predicted 3D models. Second, they only allow limited use of standard genetic variant input formats, such as Ensembl, VCF, variant identifiers (rs Id) and Human Genome Variation Society notations. The lack of support for input from a wide range of genomic-based coordinates is a major obstacle to use by the genetic-based community. To address these limitations, we developed PhyreRisk (Fig. 1), a user-friendly and publicly accessible "one-stop-shop" web application, specifically designed to bridge genomic, proteomic and structural data, and facilitate mapping of human variants onto protein structures. In addition, the PhyreRisk database providing a comprehensive resource linking human protein sequences to both experimental and Phyre-predicted structures should have applications in a wide range of studies including drug development.

PhyreRisk is available at http://phyrerisk.bc.ic.ac.uk

PhyreRisk Overview

The combination of the following key features makes PhyreRisk a valuable resource:

- User defined genetic variants can be inputted using both genomic coordinates (human built GRCh37 or GRCh38) or proteomic coordinates.
- The Protein page provides dynamic display of sequence-structure mapping onto experimental and Phyre predicted models (Fig. 1).
- Structural coverage is displayed graphically with coordinates that can be downloaded.
- Sequence and models for canonical and isoforms are provided.

PhyreRisk is implemented in Java. A guided online tutorial is available on the home page to help users navigate and learn how to use PhyreRisk.

PhyreRisk database and data sources

Figure 2 shows the data sources integrated into PhyreRisk. PhyreRisk contains fasta sequences for 20,214 human proteins, which are presented as the canonical forms based on the UniProt [16] database, and for 22,271 protein isoforms (i.e., proteins derived from alternative splicing or the use of alternative promoter or start codons, as per UniProt definition). In addition, 55,732 experimentally derived protein–protein

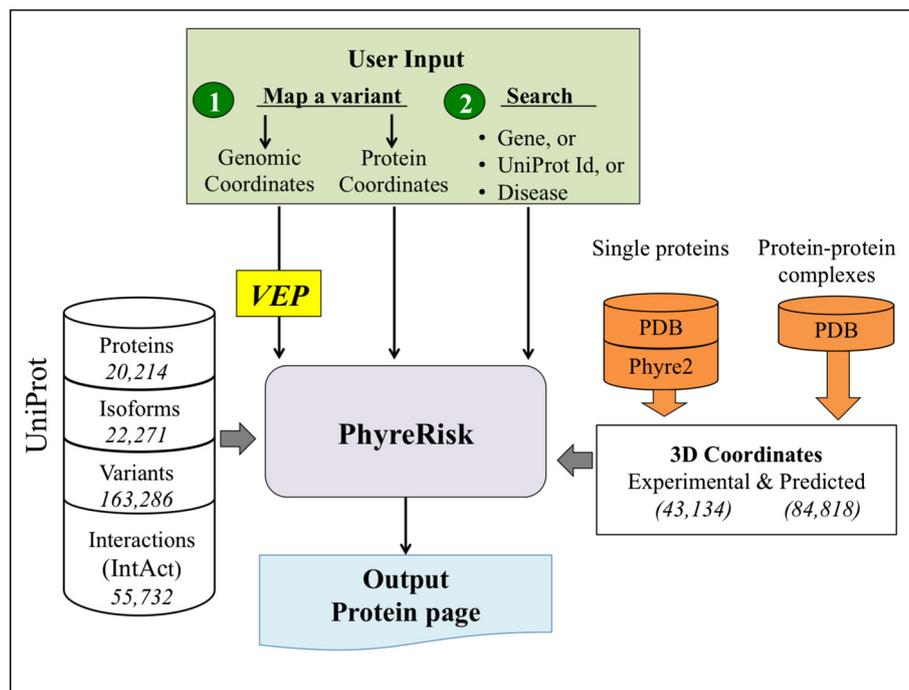**Fig. 1.** Example of Protein page displayed by PhyreRisk.

Fig. 2. Overview of the PhyreRisk pipeline.

interactions supported by multiple observations from the IntAct [17] database (as per UniProt filtering) are stored in the database. Currently, 163,286 variants from UniProt Humsavar database have been curated. In the current version of PhyreRisk, we have chosen not to store all known human variants catalogued by other databases, such as ExAC [18] and dbSNP [19].

PhyreRisk stores all its information in a postgreSQL relational database. Automatic update of the database is currently manually triggered.

Structural coverage of canonical sequences and isoforms: experimental structures and models

The greatest strength of PhyreRisk is the structural coverage of the human proteome. The database contains 18,874 experimental structures corresponding to single proteins (tertiary structures) from PDB. Moreover, it stores 84,818 pre-built predicted tertiary structures corresponding to canonical and isoform protein sequences generated using our *in house* Phyre2 software [5]. Overall, PhyreRisk provides structural coverage (partial or complete) for 14,035 (70%) out of 20,214 UniProt canonical protein sequences, of which 7987 proteins are covered by a predicted 3D model.

Because PhyreRisk aims to provide users with as much structural information as possible, especially in terms of the effect of variants on a biological system rather than just a single protein, the database incorporates all 24,260 experimental structures of protein complexes available from PDB. No selection

criteria were applied to this dataset. We are working on future development of PhyreRisk to also incorporate predicted 3D coordinates of protein complexes from GWIDD [20].

The Input Page

The PhyreRisk pipeline can handle three types of input data: (i) the user's own set of variants in genomic coordinates, (ii) the user's own set of variants using amino acid coordinates or (iii) a gene name, UniProt Id or disease name.

Genomic variants input page

Genetic variant coordinates can be described using the most commonly used formats: Ensembl, VCF, variant identifiers (rs Id) or Human Genome Variation Society notations (Fig. 3). One of the strengths of PhyreRisk is that it is a dynamic resource, which supports genome build GRCh37 and GRCh38. In contrast to the static nature of most available 3D mapping tools, such as Cosmic3D [10], which provide a list of genetic variants for which mapping is available, PhyreRisk implements a RESTful Web Service interface to programmatically query the Ensembl Variant Effect Predictor (VEP) [21]. After submitting the variant input, the PhyreRisk Results page appears within seconds and displays, among others, the variant description at protein level, its consequence term (this is the sequence ontology term
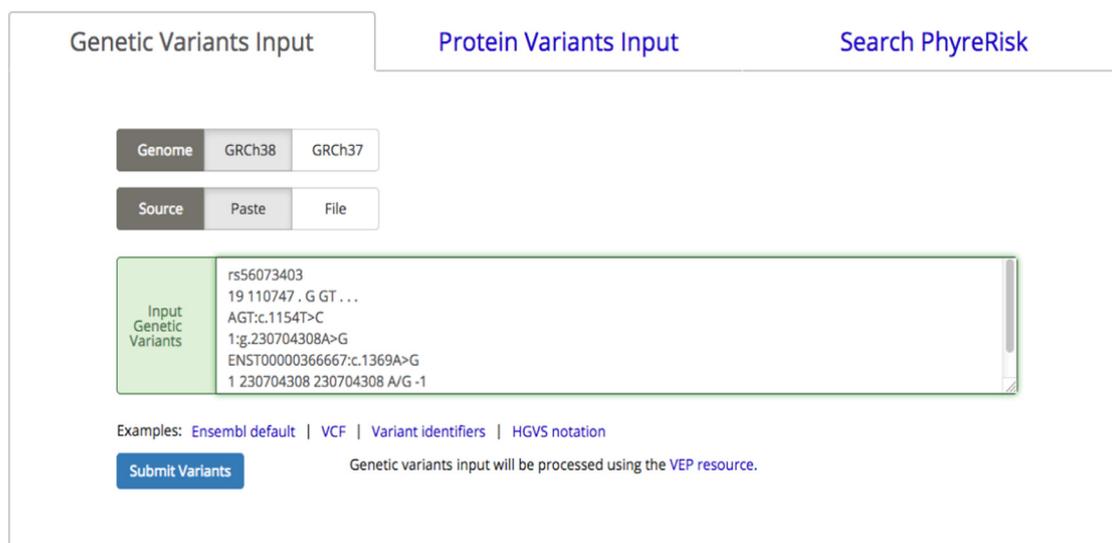
Fig. 3. PhyreRisk variant Input page.

assigned by VEP to the variant), as well as SIFT [22] and PolyPhen2 [23] *in silico* predictions (Suppl Fig. 1). All the information reported in this page is from VEP at EBI. The Results page provides a link into the PhyreRisk page, corresponding to the protein harbouring the query variants. Within the page, the amino acid under investigation is highlighted on the sequence of the protein and displayed on a 3D structure if experimental or model 3D coordinates are available.

Protein variant input page

This page can be used to explore missense variants using the amino acid position rather than their genomic coordinates (Suppl Fig. 2). The required information includes the following: the UniProt Id of the protein harbouring the substitution, the position of the wild-type amino acid, the wild-type residue and its substitution. After *submitting*, the Results page is displayed within a few seconds and provides a link to the protein's PhyreRisk page.

Search page

This page allows a search of any human protein (canonical) in PhyreRisk (Suppl Fig. 3). The search box has the autocomplete functionality enabled. PhyreRisk can be searched using different terms corresponding to the same protein. For example when searching for the low-density lipoprotein receptor adapter protein 1, one can type: (i) the gene name, for example, LDLRAP; (ii) the canonical UniProt ID, for example, Q5SW96; (iii) the UniProt entry name, for example, ARH_HUMAN; or (iv) the extended protein name, for example, low-density lipoprotein receptor adapter protein 1.

The Search page can also be used to search for "diseases." As an example, searching for "breast cancer" will return a list of all proteins, which according to the UniProt database associated with this disease.

Importantly, when searching for a protein or gene, multiple isoforms (corresponding to different transcripts) may exist. PhyreRisk adopts the UniProt classification to define the "canonical" amino acid sequence (indicated by an asterisk in PhyreRisk), and this is the one displayed by default in the protein page (see "Isoform panel" below for more information on how PhyreRisk handles isoforms).

The Protein Page

The Sequence browser

The Sequence browser named Web Alignment Viewer and Editor (WaveJS) was developed *in house* using JavaScript and WebGL. This allows fast, interactive visualization of a protein sequence with its features and annotations. It also enables partial bi-directional communication with the JSmol [24] molecular viewer. A Tooltip is available to display the structural coverage on the fasta sequence or to display variants currently available in the database or supplied by the User. The tooltip also allows choosing how to colour the 3D structure (e.g., from the default gold to cyan).

Structure selection and 3D Structure viewer panels

The structure selection panel presents in a graphical or list-view mode all the available structures (experimental or model) stored in PhyreRisk. In the graphical-view mode, the protein amino acid

sequence is presented as a bar. All available structures are also displayed as horizontal bars underneath the amino acid sequence bar, thus providing an intuitive presentation of the amino acid sequence-structure coverage. Structures are ranked and presented according to their sequence-structure coverage and structure resolution or confidence in template, for experimental structures and models, respectively.

The 3D Structure viewer panel enables the user to graphically visualize the atomic coordinate information for the selected protein structure file. Currently, PhyreRisk supports the 3D molecular viewers, JSmol, 3DMol [25], and NGL [26]. A sequence-structure mapping is performed using the built-in Structure Integration with Function, Taxonomy and Sequence (SIFTS) [27] from ePDB. However, interactive communication with the Sequence browser is so far only implemented for the JSmol viewer. It is anticipated support for bi-directional communication will be extended to NGL and 3DMol in the near future.

At present, the sequence-structure bidirectional communication allows to visualize only one residue at a time. However, the built-in JSmol viewer functions are enabled in the Structure viewer and allow easy visualization and manipulation of residues of interest on the preferred structure (examples and step-by-step guide on how to display two or more residues on the same structure are presented in the Supplementary Material). PhyreRisk is not designed to be used as a molecular viewer. For such a task, we would recommend using molecular viewers, such as Pymol (which has extensive functionality) [28] or EzMol (with guided commands designed for the occasional user) [29].

A link to our webserver Missense3D (available at http://www.sbg.bio.ic.ac.uk/~missense3d/) that provides structural analysis of the effect of a missense variant is provided (manuscript under revision in JMB).

Protein Information (summary) panel

The protein Information panel provides a high-level view of the data available for the protein of interest, such as the number of isoforms, variants, interactions and available experimental and model structures.

Isoform panel

The isoform panel presents a list of available isoforms for a given protein. The data are retrieved from UniProt and presented "as-is." Each alternative amino acid sequence is identified by its Uniprot Id, and the amino acid sequence displayed in the PhyreRisk page is highlighted. The Isoforms panel is dynamic and it allows selecting a protein isoform of interest and being redirected to the corresponding PhyreRisk page.

Variants panel and Interactions panel

The *Variants* panel presents a list of variants currently stored in PhyreRisk database for the query protein derived from UniProt database. Therefore, this is not an exhaustive list of all known variants for a given protein.

The *Interactions* panel presents a list of known experimentally derived protein–protein interactions, supported by multiple observations from the IntAct database (as per UniProt).

Documentation

PhyreRisk provides an extensive on-line tutorial available from the Home page and numerous tool tips.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.04.043.

References

[1] G. Glusman, P.W. Rose, A. Prlić, J. Dougherty, J.M. Duarte, A.S. Hoffman, G.J. Barton, E. Bendixen, T. Bergquist, C. Bock, E. Brunk, M. Buljan, S.K. Burley, B. Cai, H. Carter, J. Gao, A. Godzik, M. Heuer, M. Hicks, T. Hrabe, R. Karchin, J.K. Leman, L. Lane, D.L. Masica, S.D. Mooney, J. Moult,

G.S. Omenn, F. Pearl, V. Pejaver, S.M. Reynolds, A. Rokem, T. Schwede, S. Song, H. Tilgner, Y. Valasatava, Y. Zhang, E.W. Deutsch, Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework, Genome Med. 9 (2017), 113. https://doi.org/10.1186/s13073-017-0509-y.

[2] S.K. Burley, H.M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J.M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D.S. Goodsell, R.K. Green, V. Guranovic, D. Guzenko, B.P. Hudson, T. Kalro, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Periskova, A. Prlic, C. Randle, A. Rose, P. Rose, R. Sala, M. Sekharan, C. Shao, L. Tan, Y.-P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Woo, H. Yang, J. Young, M. Zhuravleva, C. Zardecki, RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy, Nucleic Acids Res. 47 (2019) D464–D474, https://doi.org/10.1093/nar/gky1004.

[3] S. Lyskov, F.-C. Chou, S.Ó. Conchúir, B.S. Der, K. Drew, D. Kuroda, J. Xu, B.D. Weitzner, P.D. Renfrew, P. Sripakdeevong, B. Borgo, J.J. Havranek, B. Kuhlman, T. Kortemme, R. Bonneau, J.J. Gray, R. Das, Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE), PLoS One 8 (2013), e63906. https://doi.org/10.1371/journal.pone.0063906.

[4] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The I-TASSER Suite: protein structure and function prediction, Nat. Methods 12 (2015) 7–8, https://doi.org/10.1038/nmeth.3213.

[5] L.A. Kelley, S. Mezulis, C.M. Yates, M.N. Wass, M.J.E. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis, Nat. Protoc. 10 (2015) 845–858, https://doi.org/10.1038/nprot.2015.053.

[6] J.C. Somody, S.S. MacKinnon, A. Windemuth, Structural coverage of the proteome for pharmaceutical applications, Drug Discov. Today 22 (2017) 1792–1799, https://doi.org/10.1016/j.drudis.2017.08.004.

[7] D.L. Masica, C. Douville, C. Tokheim, R. Bhattacharya, R. Kim, K. Moad, M.C. Ryan, R. Karchin, CRAVAT 4: cancer-related analysis of variants toolkit, Cancer Res. 77 (2017) e35–e38, https://doi.org/10.1158/0008-5472.CAN-17-0338.

[8] N. Niknafs, D. Kim, R. Kim, M. Diekhans, M. Ryan, P.D. Stenson, D.N. Cooper, R. Karchin, MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures, Hum. Genet. 132 (2013) 1235–1243, https://doi.org/10.1007/s00439-013-1325-0.

[9] J. Wang, R. Sheridan, S.O. Sumer, N. Schultz, D. Xu, J. Gao, G2S: a web-service for annotating genomic variants on 3D protein structures, Bioinforma. Oxf. Engl. 34 (2018) 1949–1950, https://doi.org/10.1093/bioinformatics/bty047.

[10] H.C. Jubb, H.K. Saini, M.L. Verdonk, S.A. Forbes, COSMIC-3D provides structural perspectives on cancer genetics for drug discovery, Nat. Genet. 50 (2018) 1200–1202, https://doi.org/10.1038/s41588-018-0214-9.

[11] C.J. Richardson, Q. Gao, C. Mitsopoulous, M. Zvelebil, L.H. Pearl, F.M.G. Pearl, MoKCa database—mutations of kinases in cancer, Nucleic Acids Res. 37 (2009) D824–D831, https://doi.org/10.1093/nar/gkn832.

[12] R. Mosca, A. Céol, P. Aloy, Interactome3D: adding structural details to protein networks, Nat. Methods 10 (2013) 47–53, https://doi.org/10.1038/nmeth.2289.

[13] M.J. Meyer, J.F. Beltrán, S. Liang, R. Fragoza, A. Rumack, J. Liang, X. Wei, H. Yu, Interactome INSIDER: a structural interactome browser for genomic studies,

Nat. Methods 15 (2018) 107–114, https://doi.org/10.1038/nmeth.4540.

[14] M. Ryan, M. Diekhans, S. Lien, Y. Liu, R. Karchin, LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures, Bioinforma. Oxf. Engl. 25 (2009) 1431–1432, https://doi.org/10.1093/bioinformatics/btp242.

[15] D. Wang, L. Song, V. Singh, S. Rao, L. An, S. Madhavan, SNP2Structure: a public and versatile resource for mapping and three-dimensional modeling of missense SNPs on human protein structures, Comput. Struct. Biotechnol. J. 13 (2015) 514–519, https://doi.org/10.1016/j.csbj.2015.09.002.

[16] The UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45 (2017) D158–D169, https://doi.org/10.1093/nar/gkw1099.

[17] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N.H. Campbell, G. Chavali, C. Chen, N. del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R.C. Lovering, B. Meldal, A.N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, H. Hermjakob, The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases, Nucleic Acids Res. 42 (2014) D358–D363, https://doi.org/10.1093/nar/gkt1115.

[18] K.J. Karczewski, B. Weisburd, B. Thomas, M. Solomonson, D.M. Ruderfer, D. Kavanagh, T. Hamamsy, M. Lek, K.E. Samocha, B.B. Cummings, D. Birnbaum, The Exome Aggregation Consortium, M.J. Daly, D.G. MacArthur, The ExAC browser: displaying reference data information from over 60 000 exomes, Nucleic Acids Res. 45 (2017) D840–D845, https://doi.org/10.1093/nar/gkw971.

[19] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (2001) 308–311.

[20] P.J. Kundrotas, Z. Zhu, I.A. Vakser, GWIDD: a comprehensive resource for genome-wide structural modeling of protein–protein interactions, Hum. Genomics 6 (2012) 7, https://doi.org/10.1186/1479-7364-6-7.

[21] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor, Genome Biol. 17 (2016), 122. https://doi.org/10.1186/s13059-016-0974-4.

[22] P. Kumar, S. Henikoff, P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nat. Protoc. 4 (2009) 1073–1081, https://doi.org/10.1038/nprot.2009.86.

[23] I. Adzhubei, D.M. Jordan, S.R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2, Curr. Protoc. Hum. Genet. Chapter 7 (2013)https://doi.org/10.1002/0471142905.hg0720s76 Unit7.20.

[24] Jmol: an open-source Java viewer for chemical structures in 3D, http://www.jmol.org/.

[25] N. Rego, D. Koes, 3Dmol.Js: molecular visualization with WebGL, Bioinformatics. 31 (2015) 1322–1324, https://doi.org/10.1093/bioinformatics/btu829.

[26] A.S. Rose, A.R. Bradley, Y. Valasatava, J.M. Duarte, A. Prlic, P.W. Rose, NGL viewer: web-based molecular graphics for large complexes, Bioinforma. Oxf. Engl. 34 (2018) 3755–3758, https://doi.org/10.1093/bioinformatics/bty419.

[27] J.M. Dana, A. Gutmanas, N. Tyagi, G. Qi, C. O'Donovan, M. Martin, S. Velankar, SIFTS: updated structure integration

with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins, Nucleic Acids Res. 47 (2019) D482–D489, https://doi.org/10.1093/nar/gky1114.

[28] Schrödinger, The PyMOL Molecular Graphics System, (LLC, n.d).

[29] C.R. Reynolds, S.A. Islam, M.J.E. Sternberg, EzMol: a web server wizard for the rapid visualization and image production of protein and nucleic acid structures, J. Mol. Biol. 430 (2018) 2244–2248, https://doi.org/10.1016/j.jmb.2018.01.013.